5

HANDOFF MECHANISMS TO SUPPORT REAL-TIME DELAY-CRITICAL SERVICES IN A NEXT GENERATION NETWORK

Cross Reference

10   This application claims the benefit of the Provisional application assigned U.S. serial number 60/144,597 filed on July 19 1999. Additionally, this application is related to U.S. Patent Application Docket Number 10924RRUS01P (22171.142), filed on even date herewith, entitled HANDOFF FRAMEWORK TO SUPPORT REAL-TIME DELAY-CRITICAL SERVICES IN A NEXT GENERATION

15   NETWORK, assigned to the assignee of the present application, and hereby incorporated by reference herein. Further, this application is related to U.S. Patent Application Serial Number 09/469,668, filed on December 22, 1999, entitled GPRS MAC PROCEDURES TO SUPPORT REAL-TIME SERVICES, assigned to the assignee of the present application, and hereby incorporated by reference herein..

20

Background

This application relates generally to a system and method for handing off a mobile node and more particularly to handoff mechanisms to support real-time delay-critical services in next generation wireless networks.

25   In data networks, such as Internet Protocol (IP) networks, users are typically assigned to a particular class of service (such as Platinum, Gold, or Silver) based on a service level agreement (SLA) with their service provider. In a fixed network, it is relatively easy to engineer and assign the user an appropriate amount of network resources so that the SLA can be maintained at all times. However, mobility and the

air-interface being utilized make the problem more difficult because the network resources have to be reassigned and/or renegotiated as the user moves from one cell to another. Additionally, the inherent hostile nature of the air-interface makes it difficult to predict and react accordingly to the changes in the radio frequency (RF).

5   Historically, there have been two methods to support mobility across wireless cells. In the first method, the mobile is in the full control of the decision making and the target-selection process while moving from one cell to the other. In Global Systems for Mobile Communications (GSM) or North American Time Division Multiple Access (NA TDMA) terminology, this process is known as reselection. In a

10   reselection process, the mobile node and the network have a master-slave relationship where the mobile node decides which cell serves its interest best. The network does not really have control over the target cell selection and so it is alerted to the mobile node's decision only after the target cell is selected. As a result, reselection is typically more time consuming from a network resource allocation point of view.

15   Also, reselection during an active session requires the network to temporarily buffer the data destined for the mobile node. Additionally, the target cell may not even have enough resources to address the resource needs of the mobile node.

In the second method, the network, along with input from the mobile node, decides when and where to handoff the mobile node. Handoff refers to the transfer of

20   an ongoing wireless call from one transmission site (cell) to another without disconnecting the call. This method requires constant monitoring of the mobile node's signal strength as well as complex management of target selection and network resource assignment. Network directed handoff also requires more messaging over the air compared to reselection. However, this kind of mobility

25   support can work faster because the target is known before the actual handoff takes place. Also, the resource allocation and appropriate reservation can be done at the target to meet the mobile node's demand. In a variation of this scheme in GSM and NA TDMA, the mobile node assists the network by periodically sending it RF related information regarding the mobile nodes' neighboring cells. This facilitates the

30   decision making process at the network and is called Mobile Assisted Handoff (MAHO).

The reselection based mobility works fine when the mobile node is idle or in a non-real time active session. However, as previously mentioned, reselection may result in considerable delays. Thus, buffering data for a synchronous real time application (such as voice or video) during this delay interval is not recommended.

5   Because the reselection mechanism does not have any control over the availability of the network resources at the target cell, during the few seconds of reselection related delay, several voice packets may be dropped resulting in audible speech clipping. Additionally, in current data networks, the reselection process does not prioritize resources based on the user's SLA. To address Quality of Service (QoS)

10   requirements for real-time, delay-sensitive multimedia services, the handoff mechanism needs to be optimized and enhanced in next generation wireless data (IP) networks.

Therefore, an improved system and method is desired to reduce or eliminate the aforementioned complexities and limitations.

15

Summary

In response to these and other complexities and limitations, provided herein is a unique system and method to support a handoff framework for real-time delay-critical services in a next generation wireless data network. The system includes a

20   core network (CN) and a source and a target radio network subsystem (RNS). In one embodiment, the interface between the RNSs is enhanced to reduce unnecessary relay of messaging through the CN and thus will reduce traversing delays as well as the resources utilized in message processing. Two separate methods are proposed for handing off the data traffic to the target RNS at the CN packet processing node.

25   These methods are a multi-cast function that concurrently sends data to the source RNS and the target RNS, and a suspend/resume function which suspends the data flow towards the source RNS and resumes the flow towards the target RNS.

In another embodiment, the source RNS reserves a first radio resource and detects a target RNS. The target RNS reserves a second radio resource, where an

30   amount of the second radio resource is equal to that of the first radio resource. The first RNS then performs the handoff of a mobile node to the target RNS.

Advantages are described in greater detail with respect to the drawings and the remaining disclosure.

Brief Description of the Drawings

5   Fig. 1 is a diagrammatic view of a prior art communication system.

Fig. 2 is a diagrammatic view of a communication system of the present invention for handing off a mobile node using multi-cast functionality.

Fig. 3 is a diagrammatic view of a communication system of the present invention for handing off a mobile node using suspend/resume functionality.

10   Fig. 4 is a computer of the present invention.

Detailed Description

Fig. 1 depicts a prior art communication system (network) 10 that includes a Core Network (CN) 12 that is coupled to Radio Network Subsystems (RNSs) 18-20

15   via interfaces 22. The CN 12 performs several functions including managing a user's profile while the RNSs 18-20 perform radio resource management and handoff control. The RNSs 18-20 include nodes (not shown) that transmit and receive information to and from a mobile node 14. The RNSs 18-20 are also coupled to each other via interface 16 and may pass system related information between one another.

20   According to Fig. 1, a relocation required message 24 is sent from the RNS 18 to the CN 12 when the mobile node 14 needs to perform a handoff. The CN 12 then sends a relocation request message 26 to the RNS 20. In response, the RNS 20 sends a relocation request acknowledge message 28 to the CN 12 which sends a relocation command message 30 to the RNS 18. The RNS 18 then sends a relocation commit

25   message 32 to the RNS 20 which indicates to the CN that the handoff has occurred, via a relocation detect message 34 and relocation complete message 36. Limitations of the system 10 are that it is centralized and controlled mainly by packet processing functions in the CN 12.

Fig. 2 depicts a communication system 40 of the present invention that

30   includes a CN 42 and RNSs 44-46 for performing handoff using the multicasting of messages. During the handoff process, it is important to maintain the user's QoS. As

4

such, according to the present disclosure, the following are identified as requirements during handoff in a next generation network (such as communication system 40, a 3G.IP network, a 3G.PP network, and/or a 4G network):

1. Minimizing the handoff delay for real-time, non-elastic applications (such as voice and video);

2. Initiating handoffs at the right time to effectively manage RF related issues;

3. Prioritizing handoffs based on the user's SLA and application being used; and

4. Maintaining the SLA across the handoff boundary.

Delay Minimization

A primary objective for efficient hand-off for real-time services is the reduction of delay. This can be achieved in the following ways :

- a prior resource reservation for hand-off;
- minimization of the number of messages that need to be exchanged to complete the hand-off process; and
- closely coupling the source RNS (such as RNS 44) relocation process with the movement of the user equipment (UE) such as the mobile node 14, a laptop computer or any other device that can transmit and receive data to optimize the payload traffic route through the network. It should be noted that source RNS relocation and hand-off are two separate processes.

The entire hand-off mechanism can be divided into three broad sub-functions :

1) Hand-off decision (which is not within the scope of this invention)

2) Resource reservation (for both radio and wireline)

3) Path switching. This will involve switching between nodes within the same RNS, as well as that between different RNSs.

The resource reservation and path switching functions will be discussed further.

Resource Reservation

A new Media Access Controller (MAC) state machine for wireless data systems is proposed where a new "Packet Stand-by" state has been added to support real-time, delay-sensitive services. In such a scenario, the UE 14 clings to some

5   (such as a reduced amount of) radio resource in the "Packet Stand-by" state, despite the fact that the application does not have any data to transmit. This is required to facilitate the rapid initiation of data transmission (without going through the MAC contention mechanism to acquire a new channel) as soon as the application is ready to send data.

10   In the case of handoff for real-time services, resource reservation is initiated as soon as the target node (such as RNS 46) is detected by a handoff-decision function. The amount of radio resource reservation is dictated by the MAC state of the UE 14 during handoff. A handoff controller interacts with a Radio Resource Manager (RRM) to reserve the same amount of resource in the target node that the

15   UE 14 was using in the current node (the handoff controller and the RRM are not shown but are contained in each of the RNSs 44-46). Thus the UE 14 does not have to go through the resource contention phase even though it does not have data to send.

Additionally, the resource "wastage" in the target node is minimized because resource allocation is based on the MAC state machine of the UE 14, and is not fixed.

20   A "soft" reservation mechanism is also used such that the reserved resource can be used by other low priority, non-real-time services until the user requests data to be transmitted. In that event, the lower priority traffic is pre-empted and the resource is allocated to the user. If the RRM in the target node is not able to admit the handoff due to resource constraints, then the QoS can be renegotiated through explicit

25   messaging or via an implicit policy decision (embedded in the user's SLA). For example, a user subscribing to "bronze" service can be either degraded or dropped, while a "gold" service subscriber may only be degraded, but never dropped.

Path Switching

30   When the mobile moves between nodes (or cell sites) within the same RNS, and assuming each node is connected to a Radio Network Controller (not shown) in a

tree-like fashion, the path switching is trivial and controlled by the handoff controller. In case of inter RNS hand-off, two cases might arise :

1) the mobile node14 is always anchored at a serving RNS, and depends on RNS-RNS forwarding to receive or transmit data. This handoff scenario is not associated with source RNS relocation; and

2) the mobile node 14 performs handoff with SRNS relocation.

The handoff scenario that is not associated with the source RNS relocation is unsuitable for real-time traffic because the path delay may increase proportionately with the forwarding (or tunneling) paths. This type of forwarding chain is generally not very scalable in IP networks due to the additional overhead of maintaining tunnel state information. Also, the current mechanism of performing source RNS relocation, which was described in Fig. 1, is unsuitable for real-time services. The source RNS relocation should be closely coupled with the actual hard handoff process and occur with minimal delay to reduce the "mute period" of real-time, delay-sensitive services during handoff. Thus it also makes sense to couple the source RNS relocation process with the resource reservation function (discussed earlier) for the sake of efficiency and reduction of delay.

The present invention enhances the interface 16 between the RNSs 18-20 in Fig. 1 to perform some of the messaging that is transmitted and received via the interface 22. This enhancement will reduce the unnecessary relay of the messaging through the CN 12 and thus will reduce the resources utilized in message processing and traversing delays. Two separate methods are proposed for handing off the data (or bearer) traffic to the target RNS (such as RNS 46) at the CN 42 packet processing node. These methods are a multi-cast function to the source RNS and the target RNS, which is described further in Fig. 2, and a suspend/resume function (suspending the flow towards the source RNS and resuming the flow towards the target RNS) at the CN 42, which is described further in Fig. 3.

Referring again to Fig. 2, in operation, the source RNS 44 decides on a target RNS 46 (this target RNS selection is closely coupled with the handoff threshold detection) and sends a Relocation Request message 27 to the target RNS 46 via the

enhanced interface 17. The Relocation Request message 27 is enhanced to contain details of the radio access bearer identifier, resource reservation information (as discussed above), an IP address of the packet processing node serving the UE 14 in the CN 42, as well as any uplink tunnel state information. The source RNS 44 sends

5 a Relocation Started message 48 to the packet processing node serving the UE 14 in the CN 42. Upon receiving the Relocation Started message 48, the CN 42 waits for a Relocation request ACK message to be received from the target RNS 46. The target RNS 46 reserves the radio resources and bi-casts the Relocation request ACK message 29 to the packet processing node in the CN 42 and to the source RNS 44.

10 This message 29 will contain the IP address of the packet processing function allocated to the UE 14 within the target RNS 46 and sets up the tunnel to the CN 42 for the uplink packets. The CN 42 sets up the tunnel for downlink packets (which is a duplicate copy of the tunnel towards the source RNS 44) towards the target RNS 46. The CN 42 then begins to bi-cast the downlink packets towards the source RNS 44

15 and the target RNS 46. These packets are buffered by the target RNS 46. In the meantime, the source RNS 44 handoff controller (not shown) completes the handoff process to the target cell (not shown) which is within the target RNS 46. The source RNS 44 transmits to the target RNS a Relocation Commit message 33 that contains information regarding the last packet that has been sent to the UE 14. The target RNS

20. 46 starts then starts transmitting to the UE 14 from the next packet onwards in its buffer.

In order to achieve a smooth handoff, the CN 42 (especially if it is non-QoS guaranteed) and a network between the RNS's 44-46 (such as a public IP network with no guarantee on delay) should be well-engineered to ensure that the above

25 · messages described in Fig. 2 do not take a long time to reach their destinations thus prolonging the mute period during handoff. If this engineering is not possible, a mute buffer can be used at the target RNS 46, with a mute period estimated from network delay behavior. If the mute buffer solution is utilized, the Relocation Commit message 33 may not need to contain the information regarding the last packet that

30 was sent to the UE 14.

Fig. 3 describes the suspend/resume function at the CN 42. Similar to the steps described in Fig. 2, in operation, the source RNS 44 decides on the target RNS 46 and sends a Relocation Request message 27 to the target RNS 46 via the enhanced interface 17. The Relocation Request message 27 is enhanced to contain details of

5 the radio access bearer identifier, resource reservation information (as discussed above), an IP address of the packet processing node serving the UE 14 in the CN 42, as well as any uplink tunnel state information. The source RNS 44 sends a Relocation Started message 48 to the packet processing node serving the UE 14 in the CN 42. Upon receiving the Relocation Started message 48, the CN 42 waits for the

10 Relocation request ACK message to be received from the target RNS 46. The target RNS 46 reserves the radio resources and bi-casts the Relocation request ACK message 29 to the packet processing node in the CN 42 and to the source RNS 44. This message 29 will contain the IP address of the packet processing function allocated to the UE 14 within the target RNS 46 and sets up the tunnel to the CN 42

15 for the uplink packets. The CN 42 sets up the tunnel for downlink packets towards the target RNS 46. The CN 42 then begins to bi-cast the downlink packets towards the source RNS 44 and the target RNS 46. These packets are buffered by the target RNS 46. In the meantime, the source RNS 44 handoff controller (not shown) completes the handoff process to the target cell (not shown) which is within the target

20 RNS 46. The source RNS 44 transmits to the CN 42 the Relocation Commit message 33 that contains information regarding the last packet that has been sent to the UE 14. The CN 42 suspends traffic flow towards the source RNS 44 and resumes transmission towards the target RNS 46.

In this scenario, if the network engineering to ensure that the above messages

25 described in Fig. 3 do not take a long time to reach their destinations is not possible, no mute buffer would be required at the target RNS 46.

Fig. 4 depicts a computer 60 that comprises a processor 62 and memory 64. The computer 60, which contains a computer program, may be a personal computer or laptop, the CN 42, the RNSs 44-46, the UE 14, and/or any device that can transmit

30 and receive handoff related information. The processor 62 may be a central processing unit, digital signal processor, microprocessor, microcontroller,

microcomputer, and/or any device that manipulates digital information based on programming instructions. The memory 64 may be read-only memory, random access memory, flash memory and/or any device that stores digital information. The memory 64 is coupled to the processor 62 and stores programming instructions

5   (contained in the computer program) that, when read by the processor 62, cause the processor to perform the functionality discussed above with reference to Fig.'s 2 and 3.

The present invention thus enjoys several advantages. For example, two methods are proposed for handing off data traffic to a target RNS. These methods,

10   the multi-cast function and the suspend/resume function, achieve two fundamental purposes: closely coupling the hand-off process (with resource reservation) with the source RNS relocation process to reduce hard handoff latency for real-time services, and making the source RNS relocation more efficient by reducing the amount of messaging and the involvement of the CN.

15   It is further understood that other modifications, changes and substitutions are intended in the foregoing disclosure and in some instances some features of the disclosure will be employed without corresponding use of other features. Additionally, singular discussion of items and/or computers located in the system 40 is also meant to apply to situations where multiple items and/or computers exist.

20   Further, the system 40 may include additional and/or fewer items and/or computers that perform similar functions discussed in the disclosure. Accordingly, it is appropriate that the appended claims be construed broadly and in a manner consistent with the scope of the disclosure.